ISSN 1870-4069

Suicide Tendency Content Detection with Natural Language Processing and LIME Explainer

Alberto Alvarado Sandoval[1], Fernando Aguilar-Canto[2], Diana Jiménez[2], Hiram Calvo[2]

[1] Instituto Politécnico Nacional, Escuela Superior de Cómputo, Mexico

[2] Instituto Politécnico Nacional, Centro de Investigación en Computación, Mexico

{alberto.alvarado.sandoval, pherjev, dianaljl.99, hiramcalvo}@gmail.com

Abstract. Suicide is a widespread global concern, particularly among the young population with high social media usage. Detecting suicide tendencies early can offer crucial aid. This study investigates the use of Natural Language Processing (NLP) techniques, including Language Models, to identify suicide-related content. By training a language model on a dataset containing both public content and social media posts from individuals with documented suicide tendencies, we aim to develop a tool for recognizing language patterns suggestive of potential suicide risk. Additionally, we explore the integration of the LIME explainer to enhance local interpretability, improving model comprehension. This paper provides a comprehensive exploration of identifying suicide-related text using NLP, employing diverse methodologies including classical machine learning and state-of-the-art Large Language Models (LLMs) like BERT, RoBERTa, and DistilBERT. Remarkably, the DistilBERT model surpasses more complex counterparts, acchieving 0.97741 of validation accuracy and 0.97584 of testing accuracy. Introducing an explainer algorithm improves model transparency, illuminating decision processes. Our findings emphasize the potential of advanced NLP techniques for understanding and addressing suicide-related content, benefiting mental health professionals and digital platforms striving to offer timely support and intervention.

Keywords: Suicide content detection in texts, large language models, logistic regression, random forest

1 Introduction

Suicide is a concerning global public health issue. According to the World Health Organization (WHO), approximately 700,000 people take their own lives every year, making it a leading cause of death, especially among young populations aged 15 to 29 years [23]. One crucial aspect of suicide prevention is recognizing the warning signs displayed by individuals at risk [25, 8].

These signs often manifest in their language and expressions, and with the widespread use of the Internet and social media, such content has become more accessible for analysis [6]. In recent years, Natural Language Processing (NLP) techniques, particularly Language Models, have shown remarkable capabilities in understanding and generating human-like text [3]. These Language Models have been applied to a wide range of tasks, from language translation to sentiment analysis.

Recognizing the potential for early detection of suicide tendencies in online content, we propose a research project that leverages Language Models to identify and analyze linguistic markers associated with suicide risk. The objective of this study is to investigate the viability of employing Language Models and NLP techniques for detecting suicide tendencies. By training a language model on a dataset comprising public content and social media posts from individuals with documented suicide tendencies, we aim to create a tool that effectively identifies language patterns indicative of potential suicide risk.

This tool could be integrated into social media platforms and online communities to automatically flag concerning content and offer timely support and resources to those in need. Additionally, we are considering integrating the LIME explainer to provide local explainability, aligning with one of the key directions indicated by Ji et al. (2020) [15]. In this paper, we showcase a range of machine learning models, including language models, for classifying text as either containing suicide-related content or not.

Despite the prevalence of deep learning in this domain, the utilization of language models with explainability remains limited, which is our main contribution to this topic. By highlighting the potential of machine learning models in detecting suicide-related content, our aim is to contribute to suicide prevention initiatives and foster a more supportive online environment for vulnerable individuals.

2 Related work

The detection of suicide-related content in texts is a challenging task in Natural Language Processing (NLP). Various techniques, including simple keyword detection [12, 30, 13] and sophisticated Deep Learning classifiers, have been explored by different authors. Suicide-related keywords, such as “kill,” “suicide,” and “depressed,” are associated with intense negative emotions like anxiety and hopelessness, as well as social factors like family and friends [15].

Ji et al. (2018) [16] investigated several algorithms, including Support Vector Machines, dense neural networks, random forest, XGBoost, and Long-Short Term Memory (LSTM) models, using data from Reddit SuicideWatch and Twitter. The XGBoost classifier outperformed other approaches, achieving an F1-score of 0.9583 and an AUC of 0.9569.

Other authors have also explored recurrent networks, such as Coppersmith [6], who used a bidirectional LSTM with self-attention and GloVe word embeddings. Shing et al. [28] used data from Reddit SuicideWatch to label posts according to perceived risk and implemented Convolutional Neural Networks (ConvNets) for classification. Gaur et al. [9] achieved improved results by enriching ConvNets with knowledge bases in measuring suicide risk tendency.

Tadesse et al. [29] compared Deep Learning models (LSTM and ConvNets) with classical approaches, obtaining better results with the Deep Learning techniques. Mohammadi et al. [21] combined ConvNets with different recurrent networks (LSTMs, Bidirectional LSTMs, GRU) and a Support Vector Machine meta-classifier for suicide risk assessment, excelling in subtasks A and C.

Matero et al. [20] presented a suicide risk assessment approach using Bidirectional Transformers for Language Understanding (BERT) embeddings with a neural dual-context model based on two GRU networks with attention, achieving better results in subtask B. Kodati and Ramakrishnudu [17] also used BERT embeddings with recurrent networks (LSTMs, GRUs) and ConvNets, reaching a maximum F1-score of 0.959 in their Reddit dataset.

Benton, Mitchell, and Hovy [2] employed multitask learning (MTL) to detect mental health issues, including suicide attempts, using Twitter data. Ophir et al. [22] developed Single Task and Multiple Task Models to predict suicide risk from Facebook posts, achieving better results with the Multiple Task Model. Both models utilized ELMo embeddings for word vectorization. Ji et al. (2020) [15] implemented Relation Networks with Attention in the UMD Reddit Suicidality Dataset, outperforming other Deep Learning models (ConvNets, LSTMs, Bidirectional LSTMs) in predicting risk levels with three labels indicating the level of risk.

2.1 Transformers

With the recent success of the Transformer architecture in various NLP tasks, authors have applied it to the suicide classification problem. Haque et al. [10] fine-tuned Language Models (BERT, ALBERT, RoBERTa, XLNET) and achieved better results than a Bidirectional LSTM. RoBERTa attained an F1-score of 0.9547 on Reddit data. Similarly, Ananthakrishnan et al. [1] implemented BERT, DistilBERT, ALBERT, RoBERTa, and DistilRoBERTa on a Twitter dataset, with RoBERTa yielding the best results.

Additionally, Zhang et al. (2021) [31] developed the TransformerRNN architecture, outperforming other classifiers, including Naive Bayes, ConvNets, and various LSTMs, including a Bidirectional LSTM with attention, in classifying suicide, last statements from executed prisoners, and neutral posts. Sawhney et al. [27] introduced STATENet, a time-aware transformer-based model, for identifying suicidal intent in English tweets by incorporating historical emotional context.

The model outperforms other methods (ConvNets, random forest, a BERT-based classifier), highlighting the importance of emotional and temporal cues in assessing suicide risk on social media. Burkhardt [4] evaluates the utility of social media-derived training data for suicide risk prediction in clinical settings and develops a metric for assessing the clinical usefulness of automated triage.

Their BERT-based model with multi-stage transfer learning improves suicide risk prediction, achieving a F1-score of 0.797 in a Reddit dataset. The study demonstrates the potential of leveraging social media data for better risk assessment and improved clinical outcomes in suicide prevention interventions. For languages other than English, Hassib et al. [11] focus on identifying depression and suicidal ideation in the Middle East, specifically in Egypt, where suicidal deaths are prevalent.

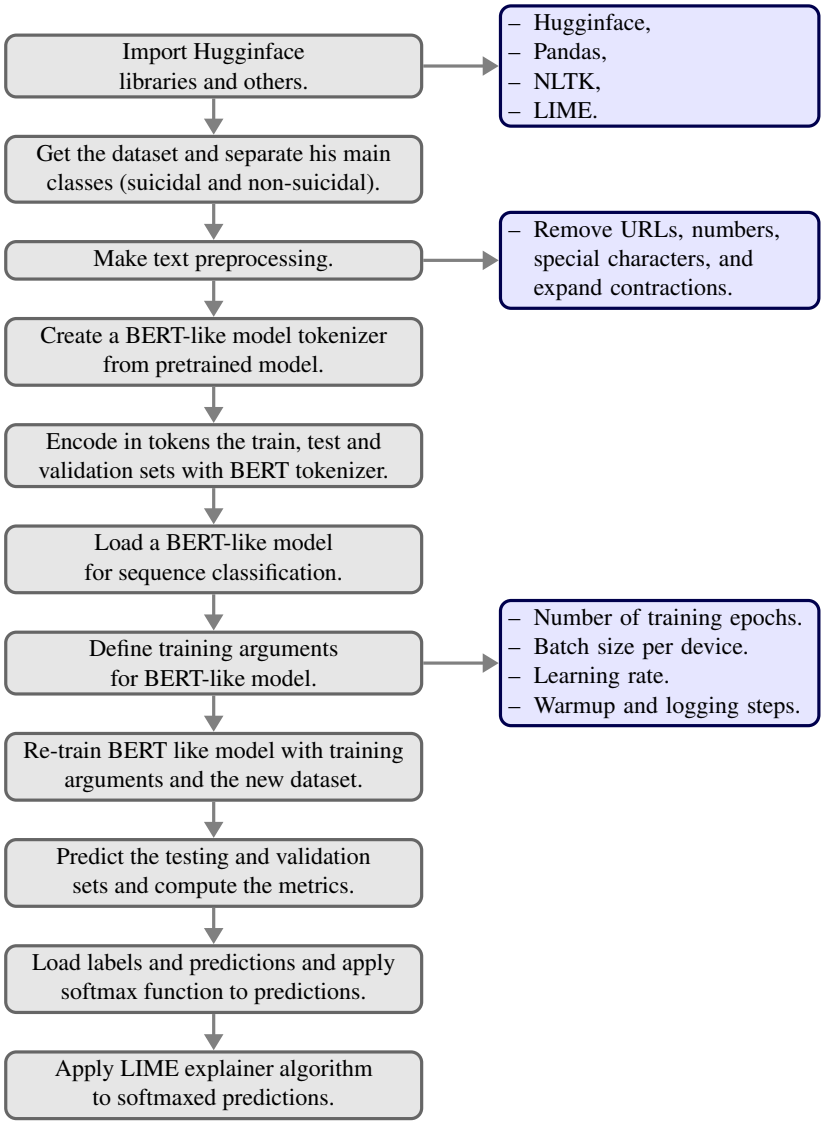

Fig. 1. Diagram of the process.

Due to a lack of mental health awareness in Arabic culture, the authors utilize social media, particularly Twitter, where users express emotions openly. They create the AraDepSu dataset with three classes (“depressed,” “suicidal,” and “neutral”) from manually labeled tweets and train it on various Transformer-based models. MARBERT performs the best, achieving high accuracy and macro-average F1-score values of 0.9120 and 0.8875, respectively. It is important to note that this study targets a different dataset and topic than the Reddit SuicideWatch revision, making direct comparisons challenging, and privacy preservation for users is a significant concern.

Table 1. Main results.

Model	Validation Accuracy	Testing Accuracy
Logistic Regression	0.92493	0.92691
Random Forest	0.90776	0.90801
BERT-base-uncased	0.97641	0.97398
RoBERTa-base	0.97544	0.81123
DistilBERT-base-uncased	**0.97751**	**0.97584**

3 Methodology

We utilized the Hugging Face transformers library for certain parts of text preprocessing and model development. All the methodology is summarized in figure 1. The text preprocessing involved several steps, including converting all text to lowercase, removing HTML tags, URLs, numbers, punctuation, and single characters. Contractions were replaced with their expanded forms, and all stopwords were removed. The texts were tokenized using three model tokenizers:

1. BERT [7],
2. DistilBERT [26],
3. RoBERTa [19].

The models were fine-tuned with one training epoch, a learning rate of 0.00005, 10 warm-up steps, and the F1-score as the training metric. Test results were predicted using each model, and the outcomes are presented below. In addition, we incorporated the Logistic Regression and Random Forest algorithms (with 100 estimators) from the Scikit-learn library to conduct a comparative analysis of its performance against that of the Large Language Models. To facilitate this evaluation, we employed the Tf-Idf Vectorizer as the feature extraction method.

3.1 Dataset description

The dataset used in this study was obtained from the "SuicideWatch" and "Depression" subreddits on Reddit, and it is available on [18]. It consists of 232,074 posts collected using the Pushift API from December 16, 2008, to January 2, 2021. The dataset is well-suited for the research because of its large size, balanced class distribution, and organic nature of the text from Reddit posts. It contains 116,037 labeled as suicidal and 116,037 as non-suicidal texts, which is one the largest NLP datasets of the topic (see [14]). 2/3 of the dataset were devoted to the training set, whereas 2/9 were used for validation and 1/9 for testing.

4 Results

The table 1 presents the main results of the study. Various models were evaluated using both validation and testing datasets. The logistic regression model achieved a high accuracy of 92.49% on the validation set and 92.69% on the testing set.

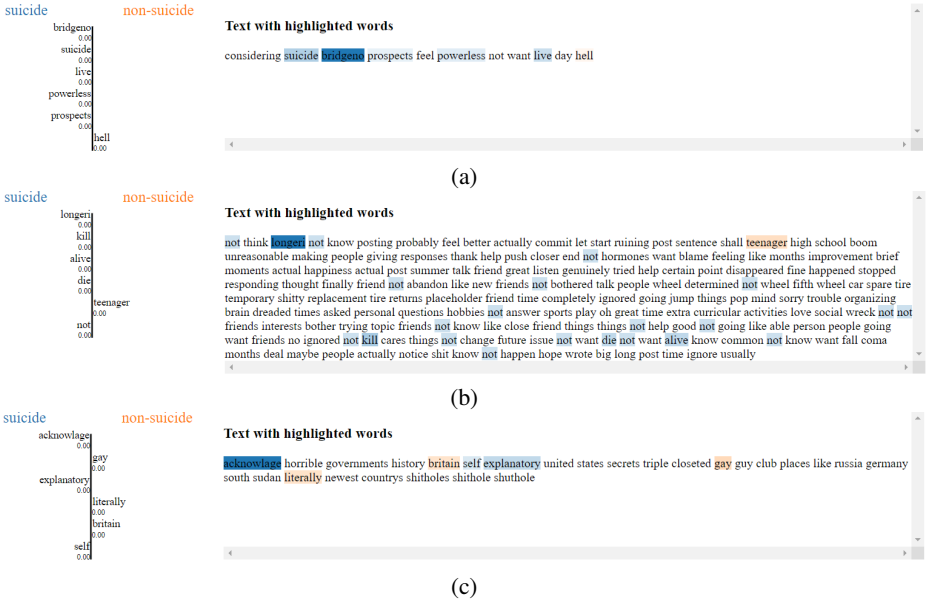

Fig. 2. Examples of the outputs of LIME with the best model (DistilBERT) with samples of the testing set.

The random forest model demonstrated slightly lower accuracy with 90.77% on validation and 90.80% on testing. Notably, BERT-base-uncased exhibited remarkable performance, achieving 97.64% accuracy on validation and 97.40% accuracy on testing. In contrast, RoBERTa-base displayed a high validation accuracy of 97.54%, but its testing accuracy dropped to 81.12%. The DistilBERT-base-uncased model outperformed the rest, boasting the highest validation accuracy of 97.75% and testing accuracy of 97.58%.

4.1 Explainability

To increase model transparency, we integrated the Local Interpretable Model-agnostic Explanations (LIME) explainer [24], which faithfully interprets predictions by locally approximating the model with an interpretable one. We gathered 6000 samples for the explainer algorithm. Figure 2 depicts the application of LIME to the optimal model, DistilBERT, while Figure 3 showcases the results of the explainer applied to the second-best model, BERT.

In the initial sentence (see Figure 2a and 4.1), the DistilBERT classifier accurately categorizes the text as “suicide.” The words “suicide,” “bridgeno,” “prospects,” “powerless,” and “live” are unsurprisingly associated with the “suicide” class. However, unexpectedly, the word ”hell” is linked to the negative category. The same results were observed in BERT. Moving on to the subsequent sentence (Figures 2b and 4.1), in the case of DistilBERT, words such as “not,” “longer,” “kill,” “alive,” and “die” exhibit a strong association with the “suicide” class, while “teenager” is associated with the non-suicidal category.

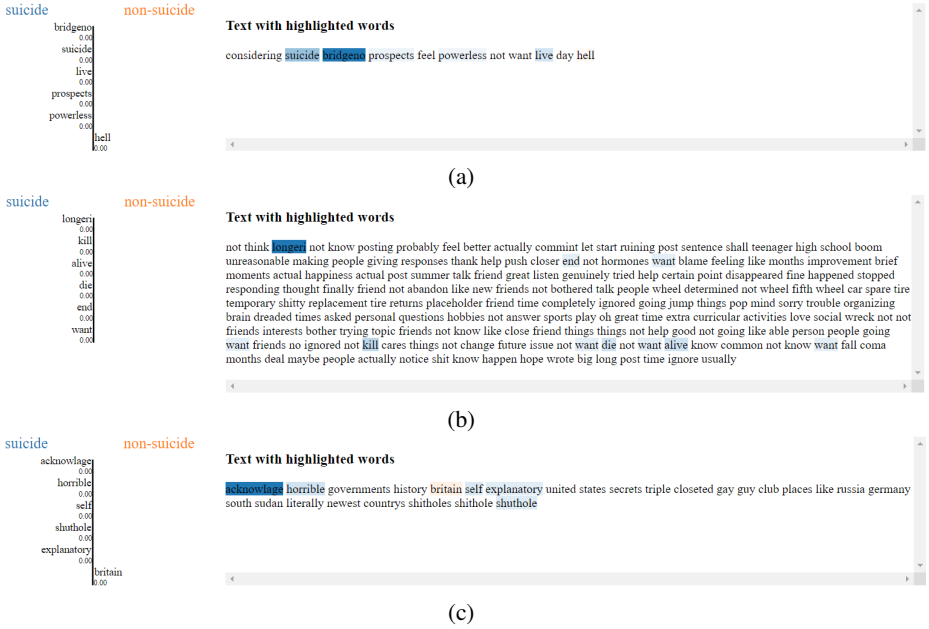

Fig. 3. Examples of the outputs of LIME with the second best model (BERT) with samples of the testing set.

A similar pattern emerges with BERT, although it does not attribute negative associations to the "suicide" class. Instead, it considers the words "end" and "wants" as associated with this class. Both models correctly classify this text as "suicidal." Finally, applying DistilBERT in the last example reveals that words like "gay," "literally," and "britain" correlate with the non-suicidal class, whereas "acknowlage," "explanatory," and "self" are linked to the "suicide" class.

Consequently, this example is correctly classified as "non-suicidal" by DistilBERT. On the other hand, the BERT model identifies "acknowlage" and "explanatory" as linked to the suicide class but adds the words "shuthole" and "horrible" to this category. It only considers the word "britain" as associated with the non-suicidal class. While this overview is not exhaustive, it provides insights into the pertinent words considered by each model, aligning with the observations made by Ji et al. (2020) [15].

5 Discussion

This study addresses the pressing need to employ modern language models in addressing the issue, echoing the call made by Ji et al. [15]. While explainer algorithms have found application in the context of depression detection [5], their utility in the suicide detection problem remains relatively unexplored. Leveraging explainers like LIME can significantly benefit professionals seeking to comprehend the decision-making processes within Deep Learning models.

Turning to the numerical results, it is important to acknowledge that while many related works have utilized data from Reddit, direct comparisons are challenging due to variations in datasets. However, as discussed in Section 2, even the most promising numerical outcomes have yielded F1-scores below 0.96. This performance metric falls short of the results achieved by DistilBERT in our study.

This suggests that the adoption of advanced language models, especially DistilBERT, holds significant promise in enhancing suicide detection capabilities, potentially offering improved support and intervention for at-risk individuals. Nonetheless, the need for continued research and exploration in this vital area is evident, particularly in expanding the utility of explainer algorithms and fine-tuning models for real-world applicability.

6 Conclusions

In this study, we have tackled the task of detecting text related to suicide tendencies, framing it as a natural language processing (NLP) classification problem. We explored the effectiveness of various machine learning approaches, including classical algorithms like Logistic Regression and Random Forest, alongside state-of-the-art Large Language Models (LLMs) such as BERT, RoBERTa, and DistilBERT. As anticipated, our experiments revealed that the deep learning-based LLMs generally outperformed the classical methods, underscoring the power of these advanced models in capturing intricate language patterns indicative of suicide-related content.

However, the unexpected outcome emerged where DistilBERT, a distilled version of BERT, exhibited superior performance compared to BERT and even RoBERTa, a larger model. This observation highlights the intricate interplay between model complexity and performance, indicating that more extensive architectures might not always lead to better results.

Building upon the insights provided by [15], we incorporated an explainer algorithm to demystify the decision-making process of our best-performing model, DistilBERT. By doing so, we aimed to enhance the interpretability of the algorithm, which could be crucial for professionals and practitioners dealing with this critical problem. The fusion of fine-tuned LLMs with explainer algorithms holds promise for the future, offering a valuable tool to understand, analyze, and address suicide tendency content more effectively.

References

1. Ananthakrishnan, G., Kumar-Jayaraman, A., Trueman, T. E., Mitra, S., Abinesh, A. K., Murugappan, A.: Suicidal intention detection in tweets using BERT-based transformers. In: International Conference on Computing, Communication, and Intelligent Systems, pp. 322–327 (2022) doi: 10.1109/icccis56430.2022.10037677
2. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 1, pp. 152–162 (2017)

3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems 33, vol. 33, pp. 1877–1901 (2020)
4. Burkhardt, H. A., Ding, X., Kerbrat, A., Comtois, K. A., Cohen, T.: From benchmark to bedside: Transfer learning from social media to patient-provider text messages for suicide risk prediction. Journal of the American Medical Informatics Association, vol. 30, no. 6, pp. 1068–1078 (2023) doi: 10.1093/jamia/ocad062
5. Byeon, H.: Advances in machine learning and explainable artificial intelligence for depression prediction. International Journal of Advanced Computer Science and Applications, vol. 14, no. 6 (2023) doi: 10.14569/ijacsa.2023.0140656
6. Coppersmith, G., Leary, R., Crutchley, P., Fine, A.: Natural language processing of social media as screening for suicide risk. Biomedical Informatics Insights, vol. 10 (2018) doi: 10.1177/1178222618792860
7. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2018)
8. Dilillo, D., Mauri, S., Mantegazza, C., Fabiano, V., Mameli, C., Zuccotti, G. V.: Suicide in pediatrics: Epidemiology, risk factors, warning signs and the role of the pediatrician in detecting them. Italian Journal of Pediatrics, vol. 41, no. 1 (2015) doi: 10.1186/s13052-015-0153-3
9. Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., Sheth, A., Welton, R., Pathak, J.: Knowledge-aware assessment of severity of suicide risk for early intervention. In: Proceedings of the World Wide Web Conference, pp. 514–525 (2019) doi: 10.1145/3308558.3313698
10. Haque, F., Nur, R. U., Al Jahan, S., Mahmud, Z., Shah, F. M.: A transformer based approach to detect suicidal ideation using pre-trained language models. In: Proceedings of the 23rd International Conference on Computer and Information Technology, pp. 1–5 (2020) doi: 10.1109/iccit51783.2020.9392692
11. Hassib, M., Hossam, N., Sameh, J., Torki, M.: AraDepSu: Detecting depression and suicidal ideation in arabic tweets using transformers. In: Proceedings of the 7th Arabic Natural Language Processing Workshop, pp. 302–311 (2022) doi: 10.18653/v1/2022.wanlp-1.28
12. Huang, Y. P., Goh, T., Liew, C. L.: Hunting suicide notes in web 2.0-preliminary findings. In: Proceedings of the 9th IEEE International Symposium on Multimedia Workshops, pp. 517–521 (2007) doi: 10.1109/ism.workshops.2007.92
13. Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., Argyle, T.: Tracking suicide risk factors through twitter in the US. Crisis, vol. 35, no. 1, pp. 51–59 (2014) doi: 10.1027/0227-5910/a000234
14. Ji, S., Li, X., Huang, Z., Cambria, E.: Suicidal ideation and mental disorder detection with attentive relation networks. Neural Computing and Applications, vol. 34, no. 13, pp. 10309–10319 (2021) doi: 10.1007/s00521-021-06208-y
15. Ji, S., Pan, S., Li, X., Cambria, E., Long, G., Huang, Z.: Suicidal ideation detection: A review of machine learning methods and applications. IEEE Transactions on Computational Social Systems, vol. 8, no. 1, pp. 214–226 (2021) doi: 10.1109/tcss.2020.3021467
16. Ji, S., Yu, C. P., Fung, S. F., Pan, S., Long, G.: Supervised learning for suicidal ideation detection in online user content. Complexity, vol. 2018, pp. 1–10 (2018) doi: 10.1155/2018/6157249

17. Kodati, D., Tene, R.: Identifying suicidal emotions on social media through transformer-based deep learning. Applied Intelligence, vol. 53, no. 10, pp. 11885–11917 (2023) doi: 10.1007/s10489-022-04060-8
18. Komati, N.: Kaggle: Competitors contributor (2021) www.kaggle.com/nikhileswarkomati
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv, (2019) doi: 10.48550/arXiv.1907.11692
20. Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., Limbachiya, P., Guntuku, S. C., Schwartz, H. A.: Suicide risk assessment with multi-level dual-context language and BERT. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology, pp. 39–44 (2019) doi: 10.18653/v1/w19-3005
21. Mohammadi, E., Amini, H., Kosseim, L.: CLaC at CLPsych 2019: Fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In: Proceedings of the 6th Workshop on Computational Linguistics and Clinical Psychology, pp. 34–38 (2019) doi: 10.18653/v1/W19-3004
22. Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., Reichart, R.: Deep neural networks detect suicide risk from textual facebook posts. Scientific Reports, vol. 10, no. 1 (2020) doi: 10.1038/s41598-020-73917-0
23. Organization, W. H.: Suicide (2021) www.who.int/news-room/fact-sheets/detail/suicide
24. Ribeiro, M. T., Singh, S., Guestrin, C.: Why should i trust you? explaining the predictions of any classifier. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97–101 (2016) doi: 10.18653/v1/N16-3020
25. Rudd, M. D.: Suicide warning signs in clinical practice. Current Psychiatry Reports, vol. 10, no. 1, pp. 87–90 (2008) doi: 10.1007/s11920-008-0015-4
26. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv, (2019) doi: 10.48550/ARXIV.1910.01108
27. Sawhney, R., Joshi, H., Gandhi, S., Shah, R.: A time-aware transformer based model for suicide ideation detection on social media. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 7685–7697 (2020) doi: 10.18653/v1/2020.emnlp-main.619
28. Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Daumé, H., Resnik, P.: Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pp. 25–36 (2018) doi: 10.18653/v1/W18-0603
29. Tadesse, M. M., Lin, H., Xu, B., Yang, L.: Detection of suicide ideation in social media forums using deep learning. Algorithms, vol. 13, no. 1, pp. 7 (2019) doi: 10.3390/a13010007
30. Varathan, K. D., Talib, N.: Suicide detection system based on twitter. In: Proceedings of the Science and Information Conference, pp. 785–788 (2014) doi: 10.1109/sai.2014.6918275
31. Zhang, T., Schoene, A. M., Ananiadou, S.: Automatic identification of suicide notes with a transformer-based deep learning model. Internet Interventions, vol. 25, pp. 100422 (2021) doi: 10.1016/j.invent.2021.100422